Single Nucleotide Polymorphisms in *Mycobacterium tuberculosis* Structural Genes

To the Editor: A recent article by Fraser et al. (1) discussed the frequency of single nucleotide polymorphisms (SNPs) in two genomes of *Mycobacterium tuberculosis*, strains H37Rv (2) and CDC1551 (unpublished). The article contains an inaccurate representation of our published *M. tuberculosis* data on SNP frequency. The authors state that "detailed comparison of strains H37Rv and CDC1551 indicates a higher frequency of polymorphism, approximately 1 in 3,000 bp, with approximately half the polymorphism [sic] occurring in the intergenic regions. In other words, 50% of the polymorphisms are in 10% of the genome. While this rate is higher than that suggested (3), it still represents a lower nucleotide diversity than found in limited comparisons from other pathogens."

On the basis of comparative sequence analysis of eight *M. tuberculosis* structural gene loci (open reading frames [orf]), we initially published an estimated average number of synonymous substitutions per synonymous site (K_s value) that indicated that this pathogen had, on average, approximately 1 synonymous difference per 10,000 synonymous sites (4). This finding was unexpected given the relatively large population size of *M. tuberculosis* and paleopathologic evidence suggesting its presence in humans as early as 3700 B.C. Subsequent sequence analysis of two megabases in 26 structural genes or loci in strains recovered globally confirmed the striking reduction of silent (synonymous) nucleotide substitutions compared with other human bacterial pathogens (3). A large study (approximately 2 Mb of comparative sequence data) of 12 genes potentially involved in ethambutol resistance (5) and 24 genes encoding protein targets of the host immune system (6) provided data consistent with the original estimate of 1 synonymous nucleotide change per 10,000 synonymous sites in structural genes in this pathogen. Our estimate did not include SNPs located in putative regulatory regions of structural genes (intergenic regions), nor did it include nonsynonymous nucleotide changes in structural genes. These classes of polymorphisms were not included in our estimates because of difficulties in ruling out the possibility that they arose as a consequence of selective pressure due to antimicrobial agent treatment or perhaps extensive in vitro passage. Synonymous nucleotide changes (neutral mutations) are commonly used to estimate many values of interest to evolutionary biologists and population geneticists.

The estimate provided by Fraser et al. is based on a genomewide frequency of SNPs (1/3,000 nucleotide sites), 50% of which presumably are located in intergenic regions and 50% in structural genes. On the basis of a genome size of roughly 4.4 Mb, there would be roughly 1,500 total SNPs, with approximately 750 in orfs (90% of genome = 3,960,000 bp) and 750 in intergenic regions (10% of genome = 440,000 bp). On the basis of these estimates, the frequency of all SNPs located in structural genes would be roughly 1/5,280 bp. (An estimate of 1,300 total SNPs [translating to 1/6,000 bp] was presented by the group at a meeting held at the Banbury Center last December.) As expected, these numbers differ from our estimate (1/10,000), in part because they contain both synonymous and nonsynonymous nucleotide polymorphisms.

We analyzed orfs (available in public databases) dispersed around the chromosome of *M. tuberculosis* strains CDC1551 and H37Rv. Surprisingly, the number of nonsynonymous SNPs exceeded the number of synonymous SNPs. We found only approximately 323 synonymous SNPs, yielding a synonymous SNP frequency of roughly 1/12,260 bp in orfs.

M. tuberculosis, a pathogen that infects one third of humans, clearly has an unusual if not unique molecular evolution history. Precise data on the frequency of its true SNPs genomewide are critical. At this point, data (3-6) are consistent with our original estimate of 1 synonymous nucleotide change per 10,000 synonymous sites in structural genes in natural populations of this pathogen.

James M. Musser
National Institutes of Health, Hamilton, Montana

References

1. Fraser CM, Eisen J, Fleischmann RD, Ketchum KA, Peterson S. Comparative genomics and understanding of microbial biology. Emerg Infect Dis 2000;6:505-12.
2. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 1998; 393:537-44.
3. Kapur V, Whittam TS, Musser JM. Is *Mycobacterium tuberculosis* 15,000 years old? J Infect Dis 1994;170:1348-9.
4. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, et al. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A 1997;94:9869-74.
5. Ramaswamy SV, Amin AG, Goksel S, Stager CE, Dou S-J, El Sahly H, et al. Molecular genetic analysis of nucleotide polymorphisms associated with ethambutol resistance in human isolates of *Mycobacterium tuberculosis*. Antimicrob Agents Chemother 2000;44:326-36.
6. Musser JM, Amin A, Ramaswamy S. Negligible genetic diversity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure. Genetics 2000;155:7-16.